Abstract Title Page
Not included in page count.

Title:

Estimating Treatment Effects via Multilevel Matching within Homogenous Groups of Clusters

Authors and Affiliations:

Peter M. Steiner, University of Wisconsin-Madison
Jee-Seon Kim, University of Wisconsin-Madison

Abstract Body

Limit 4 pages single-spaced.

Background / Context:

Despite the popularity of propensity score (PS) techniques they are not yet well studied for matching multilevel data where selection into treatment takes place among level-one units within clusters. For instance, students might self-select or get selected by parents or teachers into treatment conditions within schools. Importantly, the selection process may considerably vary across schools. For multilevel observational data with selection at level-one, two main strategies for matching level-one units (students) exist: (i) within-cluster matching where matches are only formed within clusters (schools) and (ii) across-cluster matching where treatment and control units are matched also across clusters (Arpino & Mealli, 2008; Hong & Raudenbush, 2006; Kelcey, 2009; Kim & Seltzer, 2007; Steiner, Kim & Thoemmes, 2013; Thoemmes & West, 2011). Both strategies have their own advantages and disadvantages. Within-cluster matching does not need any cluster-level covariates and, thus, the identification and estimation of causal effects relies on weaker ignorability assumptions than across-cluster matching which also requires the correct modeling of cluster-level covariates. However, within-cluster matching frequently lacks satisfactory overlap between treatment and control units. For instance, consider retaining (vs. promoting) a student as the treatment of interest. Since retention is a very extreme selection process, it is rather hard to find a comparable promoted student for each retained student within each school. However, across schools the overlap between retained and promoted students is typically better than within clusters (due to larger sample size and heterogeneity of selection across clusters). Thus, in choosing among within- and between-cluster matching one faces a bias tradeoff between the lack of overlap within clusters and the correct specification of the PS model across clusters.

In this paper we suggest a PS matching strategy that tries to avoid the disadvantages of within- and across-cluster matching. The idea is to first identify groups of clusters that are homogenous with respect to the selection model, and then to estimate the PS and treatment effect within each of the homogeneous group. This strategy has three main advantages. First, for homogeneous groups of clusters it is easier to get the PS model approximately right (the need for level-two covariates should be less important). Second, overlap within homogenous groups of clusters should be better than within-clusters. And third, because different selection process across clusters likely result in heterogeneous treatment effects one can directly investigate treatment effect heterogeneities.

Purpose / Objective / Research Question / Focus of Study:

The purpose of our study is to demonstrate that across-cluster matching within homogenous groups of clusters is less prone to bias than within-cluster matching and complete across-cluster matching (across the whole population of clusters). This is so, because across-cluster matching within homogenous groups typically relies on better overlap than within-cluster matching but on less stringent assumptions than complete across-cluster matching. In our study we investigate how one can create homogeneous groups of clusters with respect to the selection model (alternatively one could focus on the homogeneity in outcome models). If one succeeds in grouping the clusters into groups that share almost identical selection processes, the presumption

is that each group's PS model is more likely correctly specified and that the overlap between treated and control units is better than within-clusters. The group membership of clusters might be known or unknown. A manifest grouping variable might be available if teachers or administrators select students according to school-specific guidelines which one can use to derive groups of schools that share similar assignment rules. If no knowledge about the actual selection processes is available the grouping variable is latent and need to be estimated via a mixture modeling approach, for instance. Then it is interesting to investigate whether a latent grouping variable is better suited for creating homogenous groups of clusters and, thus, removes more selection bias than a manifest grouping variable. Moreover, the separate estimation of treatment effects across groups allows researchers to investigate treatment effect heterogeneities.

Significance / Novelty of study:

Though several recent studies already investigated across-cluster matching strategies they neither addressed the full complexity involved in matching units across clusters nor did they look at strategies that match within homogenous subgroups. To the best of our knowledge, this is the first study that investigates matching within homogenous groups of clusters. Given that one is able to identify a manifest grouping variable or estimate a latent grouping variable that successfully classifies clusters into groups of with homogenous selection processes, then across-cluster matching within homogenous groups should outperform within- and complete across-cluster matching in estimating the average treatment effect (ATE).

Statistical, Measurement, or Econometric Model:

Data Generating Models. In our simulation we use a model with two level-one and two level-two covariates. In order to create three different groups of clusters we used different coefficient matrices for the data-generating selection models but also the outcome models. While the heterogeneity in the outcome models is moderate (i.e., coefficients have the same sign across groups), the groups differ considerably in their selection processes (i.e., coefficients have opposite signs). For the first group of clusters, selection is positively determined by the two level-one covariates but negatively determined by the two level-two covariates. In the second group of clusters, the two level-one covariates have a negative effect on selection while the two level-two covariates have a positive effect on selection. Thus, the two selection processes are of opposite directions. Finally, the third cluster is characterized by a selection process that is only very weakly determined by the level-one covariates (here, treatment assignment almost resembles a random assignment procedure). For each of the three groups, Figure 1 shows for a single simulated data set the relation between the first level-one covariate X1 and the logit of the PS. According to the data-generating selection models, overlap within clusters, groups, and the overall population differs. Figure 2 shows for each of the three groups the distribution the level-one covariate X1 by treatment status. The plots clearly indicate that the selection mechanisms are quite different. Table 1 shows the average percentage of overlapping cases with respect to the logit of the PS. Overall, the within-cluster overlap between treatment and control cases amounts to 84% (i.e., 16% of the cases lack overlap), but across clusters the overlap is 97%. Figure 3 shows that the outcome models also vary considerably across groups (though the slopes of the level-one covariates are all positive). We also allowed for different treatment effects across groups: 5, 20, and 15 for groups 1, 2, and 3, respectively. Note that it is rather realistic for

multilevel structures to have very different selection processes but similar data-generating outcome models. While the rationales of teachers, parents, students, and peers for selecting into a treatment might strongly differ from school to school (or district to district) the data-generating outcome model are usually more robust across schools and districts.

In simulating repeated draws from the population of clusters and units, we sampled 30, 18 and 12 clusters from each of the three groups of clusters, respectively. A cluster consisted on average of 300 level-one units (sampled from a normal distribution with mean 300 and SD 50). In each of iteration of our simulation, we first estimated different PS models, then the mixture selection models in order to determine the latent group membership (assuming it is not known), and, finally, we estimated the treatment effect using different PS techniques.

PS Estimation and Matching via Inverse-Propensity Weighting. In estimating the unknown PS we used different models, some of them including cluster fixed effects. The models are estimated in for different ways: (i) within each cluster separately (for within-cluster matching), (ii) across clusters but within the three known groups (for across-cluster matching within manifest groups), (iii) across clusters but within the three estimated latent groups (for across-cluster matching within latent groups; we estimated the group membership using a mixture PS model), and (iv) across all clusters without using any grouping information (for a complete across-cluster matching). While the PS models for (i) only include the two level-one covariates as predictors, the models for (ii)-(iv) include in addition cluster-fixed effects (thus the inclusion of level-two covariates was not necessary). Given the heterogeneity of the selection models, it is clear that the PS model for (iv) does not adequately model the different selection procedures across the three groups. We used the estimated PS to derive inverse-propensity weights for ATE. We only focus on inverse-propensity weighting because our simulations but also other studies revealed that the choice of a specific PS methods does not make a significant difference.

Estimation of Treatment Effects. Since we implemented the "matching" as inverse-propensity weighting, we ran a weighted multilevel model with the treatment indicator as sole predictor. Depending on the matching strategy, we either estimated the treatment effect (i) within clusters (in this case it is a simple regression model), (ii) within the three manifest groups, (iii) within the three latent groups, and (iv) across all clusters simultaneously. Thus, analyses (i)-(iii) produced either cluster- or group-specific estimates. In order to obtain overall ATE estimates we computed the weighted average across clusters or groups, respectively (with weights based on level-one units).

Usefulness / Applicability of Method:

The findings of this study will guide researches in choosing an appropriate matching strategy for their multilevel data at hand. Particularly if selection models are heterogeneous across clusters, estimating the treatment effects within homogenous groups allows one to obtain less biased ATE estimates within and across groups. Such a matching strategy also has the advantage that it enables the investigation of heterogeneous treatment effects across groups of clusters. In order to demonstrate the usefulness and applicability of our suggested matching strategy, we apply this technique to the ECLS-K data. We investigate the effect of retaining students in Kindergarten on

first-grade reading and math outcome. (Since we are currently doing the analyses we do not yet have results for this proposal.)

Findings / Results:

The results of our simulation study are shown in Tables 2 and 3. Table 2 shows the percent of misclassified units when we derived the group membership from the estimated mixture model (with respect to the selection process). Overall, only 8% of the units were misclassified. Table 3 shows the estimated ATEs we obtained from the different matching strategies. The prima facie effects, that is, the unadjusted mean differences between treatment and control units across clusters amount to 74, -29, and 10 points for groups 1, 2, and 3, respectively (in effect sizes: 1.1, .3, and .1 SD). Given that the corresponding true effects are 5, 15, and 10 points, the selection biases within the first two groups are rather large. According to the data-generating selection model, we have a positive selection bias in the first group but a negative selection bias in the second group. There is essentially no selection bias in group 3 because selection was extremely weak. Overall, across the three groups, selection bias is still considerably large because the prima facie effect of 30 is much greater than the true effect of 9 points.

If one estimates the ATE based on a PS that has been estimated across all clusters, selection bias is removed but only a small part of it. The across-cluster estimate of 22 points is not even close to the true effect of 9 points. Though the across-cluster PS model includes the level-one covariates and cluster-fixed effects, it fails to provide a reasonable estimate of ATE because the PS model did not allow for the varying slopes across groups. Within-cluster matching overcomes this misspecification issue, but fails to provide accurate estimates for each of the three groups because of the lack of overlap within clusters. However, the overall estimate (averaged across all clusters) is 9.05 and thus very close to the true effect. But this is only a coincidence due to the simulation set up. In general, the overall estimate obtained from within-cluster matching will be biased as well (given a lack of overlap within clusters).
A better performance is achieved by across-cluster matching within known or estimated groups. If the group membership is known then the group-specific and the overall estimates are rather close to the true treatment effects. However, with the estimated group membership, the estimates are even less biased. The overall effect averaged across the three groups (8.997) is essentially identical to the true effect of 9 points. Thus, with the estimated grouping variable we achieve a less biased result than with the known grouping variable where the overall estimate amounts to 8.263 points. This is not surprising because, in estimating the group membership from the observed data, clusters that are outlying with respect to their actual group get classified into a group that better represents the outlying clusters' selection process.

Conclusions:

The results indicate that a matching strategy that first groups the data into homogeneous groups of clusters and then estimates the treatment effects via across-cluster matching within each of the groups can outperform within-cluster matching and across-cluster matching without any grouping information. In this study we demonstrated how to form homogenous groups according to the selection process. Alternatively one could also construct homogeneous groups with respect to the outcome model, or the selection and outcome model together.

Appendices

Not included in page count.

Appendix A. References

Arpino, B., & Mealli, F. (in press). The specification of the propensity score in multilevel studies. *Computational Statistics and Data Analysis*.

Hong, G., & Raudenbush, S. W. (2006). Evaluating kindergarten retention policy: A case study of causal inference for multi-level observational data. *Journal of the American Statistical Association, 101*, 901–910.

Kelcey, B. M. (2009). Improving and assessing propensity score based causal inferences in multilevel and nonlinear settings. Dissertation at The University of Michigan. Available from: http://deepblue.lib.umich.edu/bitstream/2027.42/63716/1/bkelcey_1.pdf

Kim, J., & Seltzer, M. (2007). Causal inference in multilevel settings in which selection process vary across schools. Working Paper 708, Center for the Study of Evaluation (CSE), UCLA: Los Angeles.

Steiner, P. M., Kim, J.-S., & Thoemmes, F. (2013). Matching strategies for observational multilevel data. In *JSM Proceedings*. Alexandria, VA: American Statistical Association. 5020-5032.

Thoemmes, F., & West, S. G. (2011). The use of propensity scores for nonrandomized designs with clustered data. *Multivariate Behavioral Research, 46*, 514–543.

Appendix B. Tables and Figures

Not included in page count.

Table 1. Overlap within clusters and groups (in percent of the total number of units).

	Group1	Group 2	Group 3	Overall
Overlap within groups	97.3	91.3	99.9	97.0
Overlap within clusters	85.6	72.9	98.6	84.4

Table 2. Misclassification rates (in percent).

	Group1	Group 2	Group 3	Overall
Misclassification percentage	8.2	8.1	7.2	8.0

Table 3. Treatment effect estimates by groups and overall.

	Group1	Group 2	Group 3	Overall
True treatment effects	5	15	10	9
Prima facie effect (unadjusted effect)	73.799	-28.893	9.875	30.171
Across-Cluster PS	67.363	-32.150	-0.132	22.004
Within-Cluster PS	8.333	9.555	10.065	9.051
Within-Group PS (known groups)	2.578	16.537	10.010	8.263
Within-Group PS (estimated groups)*	4.160	15.645	10.278	8.997

* For the estimated grouping variable, the true effects within the latent groups slightly differ to the ones given above.

Figure 1. Group-specific selection models with respect to the level-one covariate X1.

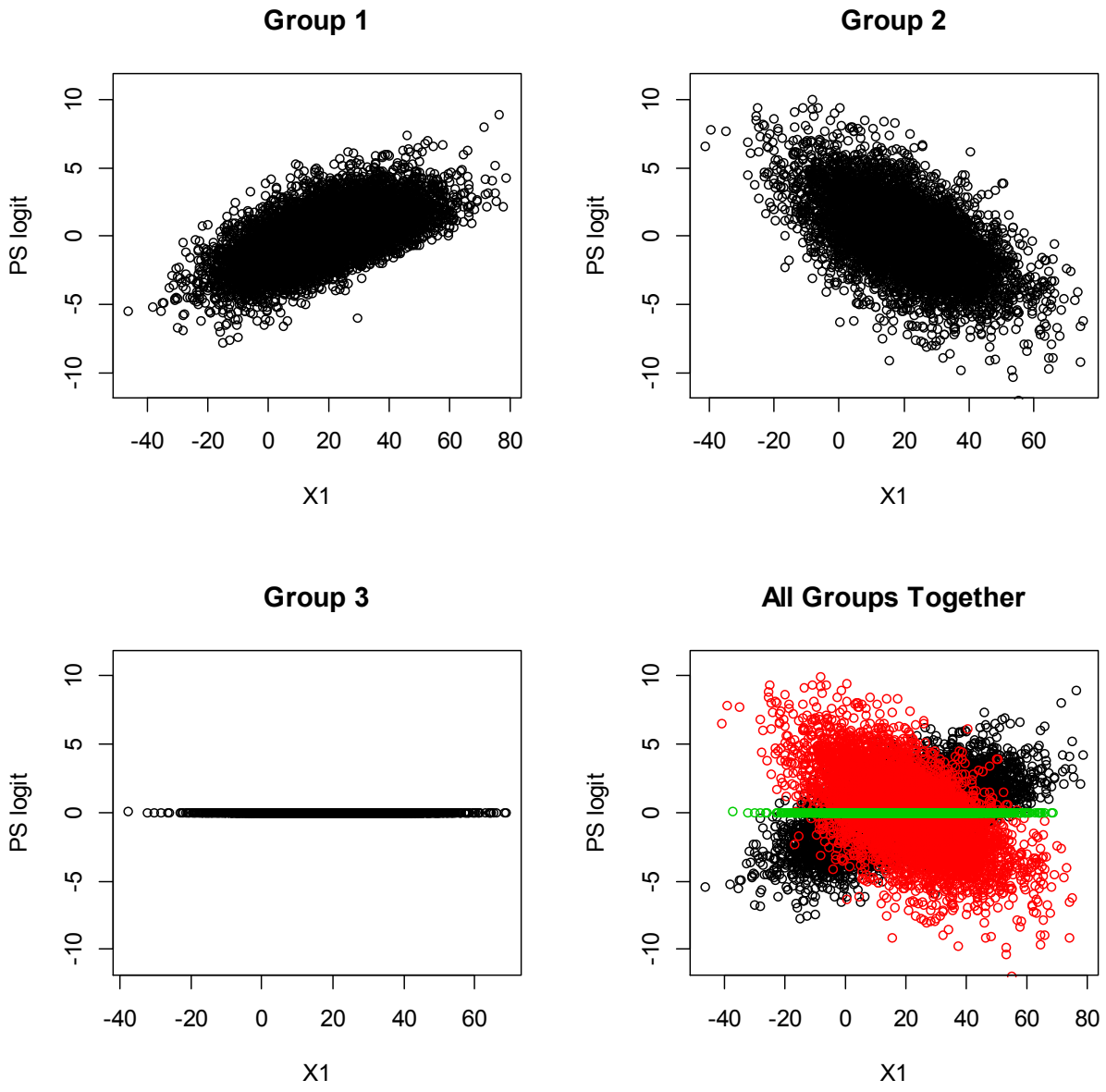

Figure 2. Distribution of level-one covariate X1 by treatment status and by group.

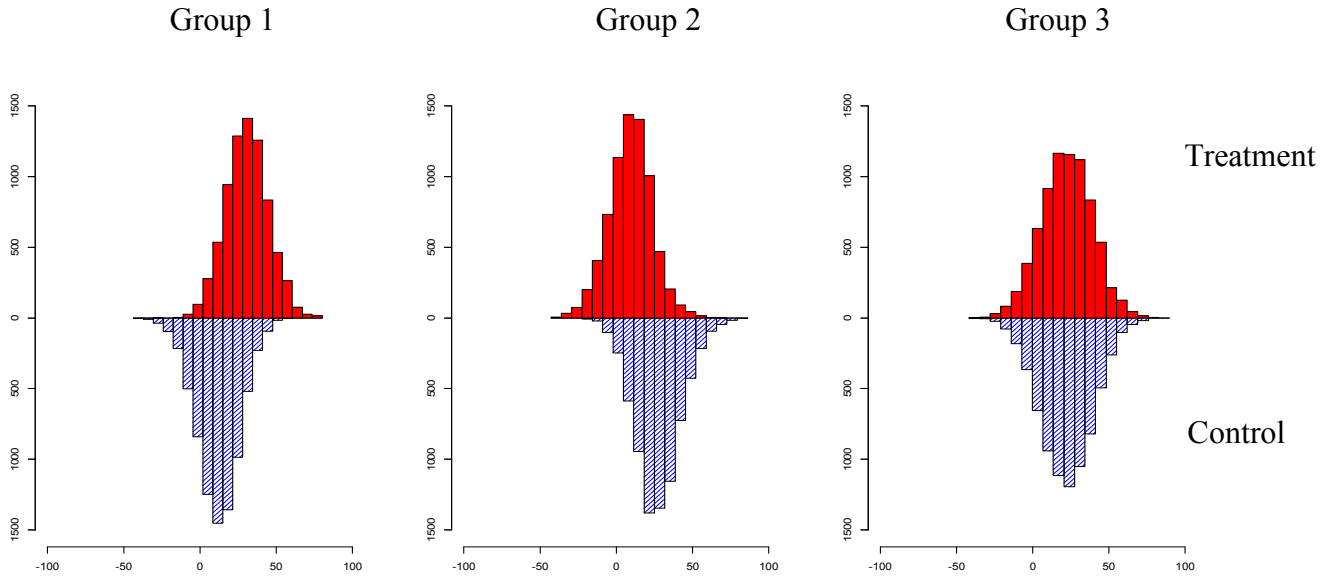

Figure 3. Group-specific outcome models with respect to the level-one covariate X1.